

Brügelmann, Hans

Kontroversen um die Schulleistungsmessung in Deutschland. Eine fiktive Diskussion über Positionen und Perspektiven in verteilten Rollen

Weinert, Franz E. [Hrsg.]: *Leistungsmessungen in Schulen*. Weinheim ; Basel : Beltz 2001, S. 33-44. - (Beltz Pädagogik)



Quellenangabe/ Reference:

Brügelmann, Hans: Kontroversen um die Schulleistungsmessung in Deutschland. Eine fiktive Diskussion über Positionen und Perspektiven in verteilten Rollen - In: Weinert, Franz E. [Hrsg.]: *Leistungsmessungen in Schulen*. Weinheim ; Basel : Beltz 2001, S. 33-44 - URN: urn:nbn:de:0111-pedocs-173952 - DOI: 10.25656/01:17395

<https://nbn-resolving.org/urn:nbn:de:0111-pedocs-173952>

<https://doi.org/10.25656/01:17395>

Nutzungsbedingungen

Gewährt wird ein nicht exklusives, nicht übertragbares, persönliches und beschränktes Recht auf Nutzung dieses Dokuments. Dieses Dokument ist ausschließlich für den persönlichen, nicht-kommerziellen Gebrauch bestimmt. Die Nutzung stellt keine Übertragung des Eigentumsrechts an diesem Dokument dar und gilt vorbehaltlich der folgenden Einschränkungen: Auf sämtlichen Kopien dieses Dokuments müssen alle Urheberrechtshinweise und sonstigen Hinweise auf gesetzlichen Schutz beibehalten werden. Sie dürfen dieses Dokument nicht in irgendeiner Weise abändern, noch dürfen Sie dieses Dokument für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, aufführen, vertreiben oder anderweitig nutzen.
Mit der Verwendung dieses Dokuments erkennen Sie die Nutzungsbedingungen an.

Terms of use

We grant a non-exclusive, non-transferable, individual and limited right to using this document.
This document is solely intended for your personal, non-commercial use. Use of this document does not include any transfer of property rights and it is conditional to the following limitations: All of the copies of this documents must retain all copyright information and other information regarding legal protection. You are not allowed to alter this document in any way, to copy it for public or commercial purposes, to exhibit the document in public, to perform, distribute or otherwise use the document in public.

By using this particular document, you accept the above-stated conditions of use.

Kontakt / Contact:

peDOCS
DIPF | Leibniz-Institut für Bildungsforschung und Bildungsinformation
Informationszentrum (IZ) Bildung
E-Mail: pedocs@dipf.de
Internet: www.pedocs.de

Digitalisiert

KAPITEL 2

Kontroversen um die Schulleistungsmessung in Deutschland

Eine fiktive Diskussion über Positionen und Perspektiven in verteilten Rollen

Hans Brügelmann

ModeratorIn: Kürzlich habe ich über die Medien folgenden Vorschlag der Krankenkassen mitbekommen: Ärzte sollten – ähnlich wie im alten China – danach bezahlt werden, wie gut und wie rasch sie Krankheiten heilen. Nur so könne die Qualität des Gesundheitswesens verbessert und der finanzielle Aufwand erfolversprechend konzentriert werden.

Dieses Beispiel wirft die Frage auf, wie eigentlich der Lehrerfolg des Schulsystems und einzelner LehrerInnen erfasst werden kann und welche Rolle dabei landesweite Testprogramme wie LAU in Hamburg, aber auch internationale Leistungsvergleiche wie die IEA-Lesestudie, wie TIMSS, PISA und PIRLS bzw. IGLU spielen können.

ArbeitgeberIn: Genau, als Steuerzahler habe ich ein Recht darauf zu wissen, was mit meinem Geld passiert und ob die Schulen ihren gesellschaftlichen Auftrag erfüllen.

LehrerIn: Ich finde diese Testerei problematisch. Die Leistungen, vor allem die fachlichen Leistungen der SchülerInnen, machen doch nur einen Teil des Auftrags der Schule aus? "Qualität" erweist sich nicht nur in kurzfristigen Ergebnissen, sondern auch in den langfristigen Wirkungen des persönlichen Umgangs der LehrerInnen mit den SchülerInnen und des inhaltlichen Anspruchs von Aufgaben und Aktivitäten und im methodischen Niveau ihrer Bearbeitung. Prozesse haben einen Eigenwert. Es ist wichtig, wie junge Menschen in unserer Gesellschaft aufwachsen sollen, und es geht damit um langfristige Wirkungen, die sich messtechnisch kaum mehr erfassen lassen. Auch den Wert der Demokratie würde doch niemand nur an ökonomischen Indikatoren festmachen wollen, oder?

Eltern: Ich verstehe gar nicht, warum in den Schulen so ein Wirbel um Tests gemacht wird. Leistungen werden doch überall bewertet. Wenn ich ein neues Auto oder auch nur einen neuen Toaster kaufen will, besorge ich mir einen aktuellen Warentest. Und auch

Personen bewertet jedeR von uns tagtäglich, z. B. wenn wir den Friseur wechseln, weil wir unzufrieden sind, oder wenn wir Bekannten unsere Hausärztin empfehlen. Warum tun sich PädagogInnen so schwer mit der Forderung nach Leistungsmessung?

LehrerIn: Tun wir doch gar nicht! Leistungsmessung gehört auch zu *unserem* Alltag. Wir LehrerInnen erfassen und bewerten Leistungen von SchülerInnen, und unsere Vorgesetzten entscheiden nach einer Unterrichtsbeurteilung über die Verbeamtung junger LehrerInnen und über die Besetzung von Funktionsstellen.

ArbeitgeberIn: Aber diese Vorgänge müssen transparenter werden! Warum reagieren Sie so empfindlich auf den Einsatz von standardisierten Instrumenten? Sie werden in Zukunft nicht mehr damit rechnen können, dass öffentliche Mittel einfach so fließen, ohne dass sich der Aufwand durch einen entsprechenden Ertrag rechtfertigen lässt. Die knappen Mittel müssen sparsam und effektiv eingesetzt werden – wie in der Industrie und wie in anderen Bereichen der öffentlichen Schulaufsicht auch. Schulen können sich zukünftig nicht mehr diffus auf “guten Unterricht” berufen, wenn ihre SchülerInnen nicht anständig lesen, schreiben und rechnen lernen.

Schulaufsicht: Sie sehen das zu technisch. Die Bewertung von Lern- und Lehrerfolgen ist aber kein Warentest, und Unterricht ist auch kein Markt, auf dem Kunden zwischen Dienstleistungen wählen – wie bei Friseur und Ärztin. Schule ist auch kein Betrieb nur zur “Produktion” von Qualifikationen, die auf dem Arbeitsmarkt verwertbar sind, sondern ein Ort der *Bildung*. Die Qualität eines Theaters oder eines Krankenhauses können Sie ja auch nicht allein nach dem Verhältnis von finanziellem Input und Output bewerten. Bildung ist mehr als eine Addition von Kenntnissen und Fertigkeiten, die man isoliert abprüfen könnte. Tests erfassen doch nur ein Oberflächenverhalten und nicht, was eine Person denkt.

LehrerIn: Genau. Denken Sie nur an die Evaluation der Programme kompensatorischer Vorschulerziehung in den 70er Jahren. Damals kamen unmittelbar nach Programmende von “Head Start” und anderen Projekten die großen Erfolgsmeldungen aus den Begleitstudien: höherer IQ, bessere Sprachfähigkeit und bessere Leistungen in den anderen trainierten Bereichen. Wenige Jahre später hieß es, die Überlegenheit der Versuchsgruppen gehe schon während der Grundschulzeit verloren, der Aufwand lohne nicht. Wieder einige Jahre später wurden sog. “sleeper”-Effekte festgestellt, d. h. diejenigen, die als Kinder an den Programmen teilgenommen hatten, erreichten bessere Schulabschlüsse, waren im Beruf erfolgreicher und hatten weniger Probleme im privaten und sozialen Umfeld.

Meine Sorge: *Kurzfristige* Leistungsmessungen erfassen wie in einer Momentaufnahme nur einen kleinen Ausschnitt der Wirkungen, der sich ohne Kenntnis der Erfahrungen im Unterricht nicht sinnvoll interpretieren, wohl aber gut als Munition im politischen Tagesgeschäft verwenden lässt.

BildungsforscherIn: Ihre Skepsis gegenüber Tests als Instrument schulischer Leistungsmessung hat eine lange Geschichte. Seit jeher wehrt sich die Schule gegen Urteile von außen und verweist auf ihre ständigen eigenen Leistungsüberprüfungen. Dabei

hätten gerade LehrerInnen allen Grund, selbstkritisch gegenüber ihrer eigenen Urteilsfähigkeit zu sein.

In den 60er und 70er Jahren wurden z. B. reihenweise Untersuchungen zur Notengebung in Klassenarbeiten publiziert. Dabei kam heraus:

- Verschiedene LehrerInnen beurteilen *dieselbe* Arbeit sehr unterschiedlich, und zwar nicht nur Aufsätze, sondern auch Rechenarbeiten und Diktate.
- Sogar *dieselben* LehrerInnen beurteilen *dieselbe* Arbeit zu verschiedenen Terminen unterschiedlich.
- Und: Urteile über (gleiche) Schülerleistungen hängen mit sozialen Faktoren wie Status der Eltern und persönlicher Beliebtheit der SchülerInnen zusammen.

Wenn man bedenkt, was alles von Noten abhängt, muss Leistungsmessung stärker objektiviert werden.

LehrerIn: Theoretisch klingt das plausibel, aber auch standardisierte Tests haben ihre Schwächen: Dieselbe Leistung kann Ergebnis angelernten Wissens oder selbstständigen Denkens sein. Und politisch ist es naiv zu glauben, in der Öffentlichkeit würden Daten aus Leistungsstudien so differenziert wahrgenommen, wie die Befunde es erfordern.

Schulaufsicht: Vor allem: Auch quantitative Daten sprechen nicht für sich selbst. Sie müssen interpretiert werden. Man müsste die Ergebnisse gleich von verschiedenen Positionen aus kommentieren, um eine einseitige Verwertung zu erschweren, wie wir sie z. B. bei TIMSS erlebt haben. So einfach ist das nicht mit der "Objektivität".

Eltern: Die ist aber für uns Eltern ganz wichtig, um die Schullaufbahn unserer Kinder verlässlich planen, und auch, um die Arbeit ihrer LehrerInnen einschätzen zu können. Insofern interessieren mich ganz konkret die Klagen der Betriebe über unzureichende, ja sinkende Leistungen der Schule. Was ist dran am sog. "Leistungsverfall"?

ArbeitgeberIn: Wir stützen uns dabei auf Ergebnisse von Tests bei den neu eingestellten Lehrlingen. Gewerkschaften, ErziehungswissenschaftlerInnen, aber auch die jeweils betroffenen Regierungsparteien stellen unsere Klagen immer gleich in Frage. Pädagogik und die Bildungspolitik verstehen sich wohl als unfehlbar. Aber ohne Bereitschaft, sich einer Außenkritik zu stellen, sind Fortschritte nicht denkbar.

BildungsforscherIn: Das sehe ich auch so. Andererseits leiden die genannten Untersuchungen unter einer Reihe von Mängeln:

- Inhaltlich erfassen die Tests nur kleine Ausschnitte schulischer Ziele (meist nur die Rechtschreibung, und die nur im Diktat, und die Grundrechenarten).
- Methodisch sind die Aufgaben oft problematisch (z. B. nicht selten exotische Wörter oder das alltagsferne und irritierende Angebot von mehreren falschen Schreibungen, unter denen die richtige auszuwählen ist).
- Für Verallgemeinerungen sind die Stichproben nicht repräsentativ.

- Für historische Vergleiche fehlen zudem vergleichbare Bezugsgruppen aus früheren Zeiten.

LehrerIn: Und Zeitungsberichte wie “Leistung zählt nicht mehr” oder gar “Neue Max-Planck-Studie widerlegt die Reformpädagogik” haben die Aussagekraft der Daten oft weit überzogen, ihre methodischen Grenzen nicht beachtet und zum Teil die Grundsätze einer fairen Interpretation empirischer Daten verletzt.

Schulaufsicht: Das ist aber doch Vergangenheit. Mit TIMSS, LAU usw. sind forschungsmethodisch versierte Institute in das Geschäft eingestiegen. Meine Sorge ist eher die Eigendynamik, die solche Testprogramme entfalten. Ob die ForscherInnen der Politik nicht doch zu viel versprechen? Um auf das Ärzte-Beispiel zurückzukommen: Wer kann eigentlich sagen, wann ein Patient “gesund” ist? Analog ist unser Problem: Kann man *Bildung* (und nicht nur Kenntnisse und Fertigkeiten) verlässlich messen? Und dann auch noch belegen, wo die Gründe für unterschiedliche Lernerfolge liegen?

ArbeitgeberIn: Wir haben doch dasselbe Problem bei der Beurteilung unserer MitarbeiterInnen in den Betrieben. Warum wollen Sie sich vor dieser Anforderung drücken? Die zentrale Frage ist doch: Ist unser Bildungssystem international konkurrenzfähig? Und eine zweite: Investieren wir in den Bildungshaushalten an den richtigen Stellen? Ein Kollege von Ihnen hat es mal treffend mit dem schönen Satz gesagt, notwendig sei eine Haltung, “die die pädagogische Version der protestantischen Rechtfertigungslehre zu überwinden trachtet, wonach es für die Erlösung allein auf den Glauben ankommt, während die Werke keine Bedeutung haben.” (Lange in VBE, 2000, S. 48)

LehrerIn: Das klingt so einfach. Als ob man nur ein Metermaß oder ein Thermometer nehmen müsste und dann daran den Zustand des Bildungswesens ablesen könnte. Aber wir haben doch schon oft die Erfahrung gemacht, dass wissenschaftliche Untersuchungen nicht zu einem klaren Ergebnis führen. Ich erinnere nur an den Methodenstreit in der Lesedidaktik, der in den 60er Jahren in einem Patt endete. Oder an die Versuche mit Gesamtschulen in den 70er Jahren, die von Verfechtern und von Gegnern bis heute unterschiedlich ausgeschlachtet werden. In den letzten Jahren ist die Integration behinderter SchülerInnen umstritten, weil die Untersuchungen keine eindeutigen Ergebnisse erbringen. Mit der “Präzision” und “Objektivität” scheint es da nicht so weit her zu sein. Ich fürchte, mit Tests erwischt man nur ein paar kleine Zipfel von dem, worum es im Unterricht wirklich geht. Schlimmer noch: In den Händen der Messtheoretiker wird aus pädagogischen Konzepten wie “Selbstständigkeit” ein mechanisches System von Fähigkeitsmodulen, in dem ich meine Idee eines partnerschaftlichen Umgangs mit Kindern und Jugendlichen nicht mehr wiedererkenne.

BildungsforscherIn: Zugegeben: Wir stecken da in einer Falle. Versuchen wir, komplexe Fähigkeiten zu erfassen, wirft man uns vor, ihr “Wesen” zu verfehlen; verzichten wir darauf, diese Fähigkeiten einzubeziehen, wird uns vorgehalten, wesentliche Ziele der Schule nicht zu berücksichtigen. JedeR von uns muss pragmatische Lösungen finden. Und wenn ich mir anschau, mit was für primitiven Aufgaben immer noch viele Lehre-

rInnen Leistungen überprüfen, dann wünschte ich mir, dass das von uns in der Testentwicklung erreichte Niveau auch nur halbwegs Standard im Unterrichtsaltag wäre.

Eltern: Ich stimme Ihnen zu. Vom grünen Tisch aus kann man leicht immer neue Forderungen erheben, was ein "vernünftiges" Messinstrument alles zu leisten habe. Alles, was besser ist als der bisherige Schlendrian, ist einen Versuch wert. Heute können LehrerInnen doch machen, was sie wollen. Und die faulen und unfähigen KollegInnen bringen den ganzen Berufsstand in Verruf. Ich verstehe nicht, dass kompetente und engagierte LehrerInnen diese Missstände decken. Sie sollten doch froh sein, wenn ihre Leistung anerkannt wird.

Schulaufsicht: So einfach ist das nicht. Vor allem hinkt der Vergleich mit den Prüfaufgaben der LehrerInnen. Landesweite Testprogramme haben eine ganz andere Funktion. Und die Nebenwirkungen falsch gewählter oder falsch verstandener Aufgaben können fatal sein. Testergebnisse haben eine hohe Suggestivkraft. Durch ihre scheinbar technische Präzision gewinnen sie das Image einer Unparteilichkeit, die in der öffentlichen Diskussion leicht überschätzt wird. Laien, und dazu zählen auch Bildungspolitikern, VerbandsvertreterInnen und JournalistInnen, können im Einzelfall die Annahmen, unter denen Aussagen von Tests gelten, und ihre konkrete Leistungsfähigkeit kaum einschätzen. Sie haben eine zu simple Alternative konstruiert: "objektive" Tests hier gegen "subjektive" (Vor-)Urteile da.

LehrerIn: Zusätzlich sehe ich Probleme in der Passung von Test und Unterricht. Vor allem die Beteiligung an internationalen Vergleichen erzwingt oft problematische Kompromisse. Da wird der Lehrplan *eines* Bundeslandes (von 16!) und der Aufgabenpool *eines* Schulbuchs zum Maßstab für "das deutsche Curriculum". Und der Zeitdruck, unter dem die Testaufgaben entwickelt worden sind, erschwert es, solchen Schwächen auf die Spur zu kommen oder gar bessere Alternativen zu entwickeln. Wir haben ja schon bei TIMSS gesehen, dass eine Reihe von Aufgaben inhaltlich oder sprachlich problematisch war.

BildungsforscherIn: Es ist sicher richtig, dass Papier-und-Bleistift- Aufgaben im Bereich Lesen und Schreiben näher an den untersuchten Fähigkeiten sind als etwa in den Naturwissenschaften. Aber in Zusatzstudien haben sich die Antworten in den Auswahlfragen der Tests als gute Annäherung an die Leistungen in Experimenten erwiesen. Die Tests sollen ja nur als *Indikatoren* dienen, sie müssen nicht die angezielten Fähigkeiten selbst erfassen.

LehrerIn: In der Theorie mag das stimmen. Aber sie müssen wieder an die Wahrnehmung durch die Betroffenen denken. SchülerInnen oder LehrerInnen, die in Tests gut abschneiden wollen, werden sich an den Indikatoren orientieren. Was bringt es z. B. noch, im Unterricht zeitlich aufwendige Experimente durchzuführen, wenn SchülerInnen Tests auch bestehen können, indem man einfach solche Testaufgaben mit ihnen übt? Wir kennen das ja vom privaten Repetitor, den StudentInnen vor dem juristischen Staatsexamen besuchen, um Prüfungsaufgaben zu pauken, statt zu studieren. "Teaching to the test" lässt sich kaum vermeiden, wenn von den Ergebnissen der Tests persönliche Vor- und Nachteile für die Betroffenen abhängen. Denken Sie nur an die Nebenwirkun-

gen von erfolgsabhängigen Verträgen für VertreterInnen in der Wirtschaft. Da wird auch ohne Rücksicht auf langfristige Folgen für die Firma oder gar für die Kunden auf rasche Erfolge gesetzt, um die Provision zu kassieren.

BildungsforscherIn: Ich finde, hier gehen verschiedene Dinge durcheinander. Wir müssen uns über den konkreten Zweck der Leistungsmessungen verständigen. Ich sehe da drei ganz verschiedene Funktionen: Landesweite Testprogramme können einmal dazu dienen, die Vergabe von Berechtigungen auf einen einheitlichen Maßstab zu beziehen. Das Zentralabitur in den süddeutschen Ländern, aber beispielsweise auch in Frankreich, hat diese Funktion. Ein anderes Ziel ist es, die Arbeit einzelner Schulen oder gar LehrerInnen zu bewerten. Vor allem in den angelsächsischen Ländern werden standardisierte Tests zu diesem Zweck eingesetzt. Bei Veröffentlichung der Daten kommt es dann zu dem viel diskutierten "ranking", einer Rangliste von Schulen. In einigen Kantonen der Schweiz wird auch überlegt, die Bezahlung der LehrerInnen an Ergebnisse in solchen Tests zu knüpfen. Eine ganz andere Funktion haben schließlich die aktuell diskutierten länderübergreifenden Stichprobenuntersuchungen, die lediglich auf einen Vergleich von Schulsystemen zielen.

Eltern: Wieso das? Wir Eltern haben doch ein berechtigtes Interesse an Informationen über die Qualität des Unterrichts in der Klasse unseres Kindes.

BildungsforscherIn: Wir müssen aktuell zwei Typen von Untersuchungen auseinanderhalten. Die KMK hat in ihren Konstanzer Beschlüssen vom Oktober 1997 einerseits länderübergreifende Untersuchungen wie TIMSS und PISA oder jetzt IGLU für die Grundschulen gefordert. Da geht es gar nicht um einzelne LehrerInnen und Schulen, insofern greift der Vorwurf des Ranking hier nicht. Der zweite Ansatz zielt auf "qualitätsverbessernde Maßnahmen" innerhalb der einzelnen Bundesländer, und da gibt es unterschiedliche Ansätze. Landesweite Testprogramme wie LAU in Hamburg sind da die Ausnahme. Und auch dort werden keine Ranglisten veröffentlicht. Schulbezogene Daten erhalten nur die Schulen selbst, als Anstoß, intern über ihre Arbeit im Vergleich zu anderen Schulen nachzudenken. Wo liegt da die Gefahr?

LehrerIn: Diese Unterscheidung kann man doch nur auf dem Papier so klar durchhalten. Wenn bei einem Systemvergleich herauskommt, dass bestimmte Schulformen – sagen wir mal Gymnasien mit Gruppierung nach Leistungsniveaus in den Hauptfächern – nach bestimmten Kriterien schlechter abschneiden, dann hat das Auswirkungen auf die Wahrnehmung der Eltern in Schulen, die nach diesem Ansatz arbeiten, und damit indirekt auch auf die Arbeit in diesen Schulen.

Eltern: Aber das ist doch gut! Wo sollen die denn sonst die Information herbekommen, was dieses System taugt?

Schulaufsicht: Na, so einfach ist das auch nicht. Wenn bei einer Studie wie TIMSS im *internationalen* Teil herauskommt dass Schulsysteme mit integrierten Sekundarschulen vergleichsweise gut abschneiden, im *Deutschland-internen* Ländervergleich aber, dass Gesamtschulen eher schlechter sind als dreigliedrige Systeme – was folgt daraus? Sol-

len wir die Gesamtschule flächendeckend einführen, weil sie ihre Qualitäten nur voll entfalten kann, wenn Gymnasien ihnen nicht die besten SchülerInnen abziehen? Müssen wir die deutschen Gesamtschulen besser ausstatten, ihre LehrerInnen besser aus- und fortbilden, damit sie das Potential der integrierten Schulform besser ausnutzen können? Oder sollten wir die Gesamtschulen ganz abschaffen, weil sie in unser gesellschaftliches Umfeld nicht so gut passen wie anderswo?

BildungsforscherIn: Möglicherweise werden hier die falschen Fragen gestellt. Die Organisationsform von Schule ist nach den vorliegenden Studien nicht der entscheidende Faktor für die Qualität des Unterrichts. Umso wichtiger ist es, dass wir die Wirkungen des *Unterrichts* erfassen – und dass wir unterschiedliche Leistungen auf konkrete Bedingungen beziehen können, um diese dann zu verbessern.

LehrerIn: Ich habe Probleme mit der Diskussion. Mal wird behauptet, die Programme zielten nur auf die Ebene des Gesamtsystems. Da frage ich mich, ob die BildungspolitikerInnen überhaupt in der Lage sind, Konsequenzen aus den Befunden zu ziehen. Was machen die z. B., wenn sich ein Ergebnis aus neueren amerikanischen Studien bestätigt, nämlich dass kleinere Klassen am Schulanfang förderlich sind und vor allem leistungsschwachen SchülerInnen zugute kommen? Werden die dann wirklich mehr LehrerInnen einstellen?

Andererseits wird gesagt, die Testprogramme sollten den Schulen helfen, ihre Arbeit vor Ort zu verbessern. Aber dann gibt es Konflikte mit dem Datenschutz und wegen der Anonymität und fehlenden Dichte der Daten sind diese dann nicht wirklich hilfreich.

Eltern: Dazu habe auch ich noch eine Frage. Landesweite Testprogramme werden ja immer wieder mit dem Argument begründet, man müsse die Arbeit der Schulen besser erfassen, um ihre Qualität steigern zu können. Warum schneidet dann aber ein Land, in dem solche Programme schon lange an der Tagesordnung sind, nämlich die USA, in internationalen Studien nicht besser ab als z. B. Deutschland? Dort mehren sich meines Wissens sogar die Stimmen, die eher eine schulnahe Aufsicht und Beratung der Schulen fordern – wie wir sie schon lange haben. Drehen wir uns da nicht im Kreis?

LehrerIn: Da möchte ich mich gleich anhängen: Bei TIMSS hat sich doch gezeigt, dass Länder mit zentralen Prüfungen nicht besser abschneiden als Bildungssysteme ohne. Woher rührt die Sicherheit, Tests seien besser als persönliche Beurteilungen und externe Leistungsmessungen führten zu mehr Qualität als schulinterne Prüfungen?

Schulaufsicht: Man muss doch sehen, dass kein System ohne Kontrolle auskommt. Manche Länder kontrollieren den "input" stärker, z. B. die Qualifikation von LehrerInnen oder die Vorgaben der Lehrpläne (so die meisten kontinentaleuropäischen Staaten). Andere geben in diesem Bereich mehr Freiräume und kontrollieren dafür den "output" stärker, z. B. durch zentrale Prüfungen (so traditionell die angelsächsischen Staaten). Das deutsche System hat sich in den letzten Jahren in Richtung auf mehr Schulautonomie bewegt, d. h. aber auch, die Aufsicht über "input" und Arbeitsabläufe verringert. Also müssen wir Rechenschaft und Kontrolle an anderer Stelle verstärken.

LehrerIn: Und dazu greifen Sie auf Instrumente zurück, deren Wert und Auswirkungen äußerst umstritten sind, ganz zu schweigen von dem Problem, dass nicht einmal die Kriterien für eine grundlegende Bildung klar sind.

ArbeitgeberIn: Über Ziele wie Lesen, Schreiben und Rechnen dürfte es aber doch keinen Streit geben.

Eltern: Naja, die Frage ist doch, welches Niveau in diesen Bereichen angesetzt wird: Was soll als Ziel für das Ende der Pflichtschule oder der Grundschulzeit gelten? Geht es z. B. um sinnbetontes Vorlesen einer Geschichte oder um die Beantwortung von Auswahlfragen zu einem Sachtext?

ArbeitgeberIn: Zu den Anforderungen kann man z. B. unsere Betriebe befragen, welche Grundfähigkeiten und welches Niveau im Berufsalltag erforderlich sind.

Eltern: Meines Wissens gibt es in dieser Hinsicht erhebliche Unterschiede zwischen verschiedenen Branchen, aber auch zwischen Betrieben innerhalb einer Branche.

LehrerIn: Zudem setzen die staatlichen Vorgaben keine klaren Prioritäten. Während jeder Lehrplan seine fachlichen Ansprüche favorisiert, stehen in den allgemeinen Richtlinien fachunabhängige Schlüsselqualifikationen wie Kooperationsfähigkeit oder Selbstständigkeit im Vordergrund. Für LehrerInnen bedeutet das, sie müssen ihre begrenzten Ressourcen auf konkurrierende Ziele verteilen. Da sind unterschiedliche Kompromisse denkbar und legitim. Die Testprogramme setzen ihrerseits eigene Prioritäten. Ich befürchte als Folge eine Uniformierung des Unterrichts, die nicht mehr auf die Besonderheiten vor Ort Rücksicht nehmen kann.

BildungsforscherIn: Ich muss noch einmal darauf hinweisen, dass es bei LAU, PISA usw. nicht um die Arbeit der einzelnen Schulen geht. Eine solche übergreifende Bestandsaufnahme kann aber erhellen, wo ein System insgesamt seine Stärken und seine Schwächen hat. Da geht es nicht um LehrerIn A oder LehrerIn B.

Schulaufsicht: Indirekt doch. Und ich denke, das ist auch so gewollt. Durch eine solche Bestandsaufnahme wird ja ein öffentlicher Druck erzeugt. Die Begriffe "Stärken" und "Schwächen" machen das deutlich. Mit den Aufgaben werden bestimmte Maßstäbe gesetzt. Die geprüften Leistungen markieren, welche inhaltlichen Bereiche besonders wichtig sind, also auch, welches Niveau in diesen Bereichen erwartet wird. Es muss deutlich werden, dass dies eine politische Frage ist. Solche Entscheidungen kann man nicht WissenschaftlerInnen überlassen.

LehrerIn: Meine Sorge ist, dass grundlegende Anforderungen nicht nur für den Schulabschluss, sondern auch für Zwischenschritte festgelegt werden. In England gibt es ja einen nach Jahrgängen gestuften "nationalen Lehrplan". Nehmen Sie den aktuellen Fall der auch in Deutschland bekannten Reformschule "Summerhill": Die SchülerInnen haben dort die Freiheit, wann sie was lernen, sie würden also bei Zwischentests eher schwach abschneiden. Am Ende der Schulzeit aber gibt es externe Prüfungen, für die

müssen sie arbeiten, und dann erreichen sie überdurchschnittliche Ergebnisse! Pädagogische Konzepte, die unterschiedliche Lernwege fördern und das individuelle Lerntempo respektieren, werden durch ein so gestuftes System gleich geschaltet. Und wenn sie dann noch an die unterschiedlichen Lernvoraussetzungen der SchülerInnen denken...

BildungsforscherIn: Wenn Sie sich Studien wie PISA, IGLU oder LAU ansehen, gilt dieser Einwand nicht mehr. Dort werden auch sozialstrukturelle Daten erhoben, um den Einfluss unterschiedlicher Milieus einzuschätzen.

Eltern: Aber selbst dieselbe Schule hat oft unterschiedliche Einzugsgebiete. So können zwei Parallelklassen in derselben Schule ganz unterschiedlich günstige Voraussetzungen haben. Auch aus demselben Einzugsbereich berichten LehrerInnen für aufeinander folgende Jahrgänge ganz unterschiedliche Zusammensetzungen.

BildungsforscherIn: Das Problem löst sich auf, wenn wir Längsschnitte anlegen, also dieselben SchülerInnen über mehrere Jahre begleiten, wie bei der LAU in Hamburg in der 5., in der 7. und in der 9. Klasse. Da vergleichen wir nicht nur die Leistung zu einem Punkt X, sondern messen Lernzuwächse.

Eltern: Aber ein Test kann doch nie umfassend die Fähigkeiten einer Person messen, sondern nur ein Verhalten in einer bestimmten Situation, und das kann wechseln.

BildungsforscherIn: Ihr Einwand ist grundsätzlich richtig. Darum werden ja auch Vertrauensintervalle angegeben, um mögliche Schwankungen bei der Interpretation zu bedenken. Man muss insofern zwischen Einzelfalldiagnose und Aussagen über Stichproben unterscheiden. Der Einfluss von Tagesform, von Missverständnissen einzelner Aufgaben usw. gleichen sich in größeren Gruppen aus. Es ist also etwas Anderes, ob wir die Leistung einzelner SchülerInnen oder die Leistung von *Gruppen* von SchülerInnen, z. B. eines Bundeslandes oder einer Schulform, miteinander vergleichen.

Schulaufsicht: Hier haben die Verantwortlichen ja auch dazugelernt. Wie schon gesagt: Die Daten werden nicht genutzt, um Urteile über einzelne Schulen, LehrerInnen oder gar SchülerInnen zu fällen.

ModeratorIn: Lassen Sie uns eine Zusammenfassung versuchen. Unsere Diskussion hat gezeigt, dass es sich bei Leistungsvergleichen um ein komplexes Problem handelt. Wir müssen m. E. drei Ebenen unterscheiden:

- die *politische*: Wer bestimmt die Kriterien, wer kontrolliert die Umsetzung -zentrale oder dezentrale Entscheidungsträger?
- die *forschungsmethodische*: Welches Instrumentarium ist das aussagekräftigste und verlässlichste – die standardisierte Messung oder das personengebundene Urteil?
- die *reformstrategische*: Auf welchen Wegen ist Schulentwicklung am ehesten in Gang zu bringen – durch eine interne oder durch eine externe Evaluation?

Außerdem gibt es je nach Standpunkt unterschiedliche Einschätzungen der Funktion und der Wirkungen des Instruments "Leistungstest". Können Sie dazu die Ihnen wichtigsten Punkte noch einmal kurz auf den Nenner bringen?

ArbeitgeberIn: Aus meiner Sicht sollten Leistungsmessungen in allen drei Funktionen eingesetzt werden:

- zur Beurteilung einzelner *SchülerInnen*,
 - um Fortschritte, aber auch Schwierigkeiten genauer erkennen und entsprechend gezielt fördern zu können und
 - um die formelle Berechtigung für den Besuch weiterführender Schulen vergleichbar und nachprüfbar zu erfassen und auszuweisen;
- zur Bewertung des LehrerInserfolgs von *LehrerInnen*, indem die Wirkungen verschiedener methodischer Ansätze unter vergleichbaren Bedingungen überprüft werden;
- zur Einschätzung von Stärken und Schwächen des *Systems* insgesamt bzw. einzelner Teilsysteme, z. B. Schularten.

Eltern: Ich wünsche mir, dass das Hintenherumreden über "gute" und "schlechte" LehrerInnen durch verlässliche Daten ersetzt wird. Aber dazu müssten zunächst einmal die Ziele der Schule in Abstimmung aller beteiligten Gruppen geklärt und klarer umrissen werden: Wo liegen die Prioritäten für Allgemeinbildung heute? Wenn die Kriterien geklärt sind, dann könnten auch Zeugnisse aus verschiedenen Schulen und Bundesländern durch zentrale Tests vergleichbarer werden. Meine Sorge: Ich glaube nicht daran, dass eine genauere Information über den "Stand" des deutschen Bildungssystems politisch etwas bewirkt. Ein Beispiel: Seit Jahren berichtet die OECD, dass Deutschland für die Grundschule, in der doch die Grundlagen für alles Weitere gelegt werden sollen, viel weniger ausgibt als andere Länder, für die Sekundarstufe dagegen mehr. Geändert hat sich bis heute nichts.

LehrerIn: Ich glaube auch nicht an die Objektivität von Tests: Der eine löst die Aufgabe, weil er kurzfristig viel auswendig gelernt hat, der andere ist durch eigenes Denken zu der Lösung gekommen. Das sind verschiedene Qualitäten.

Aber auch für die Schulen liegt mir daran, dass die Ministerien jetzt nicht nur auf das Pferd "zentrale Leistungsmessung" setzen. Was nutzt es, wenn wir wissen, dass wir schlecht sind – oder jedenfalls schlechter als andere? Wie es heute so schön heißt: Vom wiederholten Wiegen wird die Sau auch nicht fetter...

Und eine schärfere Kontrolle mag helfen, einzelne Extremfälle zu verhindern, *besser* wird der Unterricht in der Breite dadurch nicht. Um die Qualität von Unterricht zu verbessern, wäre es viel wichtiger, Mittel und Unterstützung für schulinterne Bestandsaufnahmen und vor allem für die Verbesserung der alltäglichen Arbeit und ihrer Bedingungen bereit zu stellen. Und wir brauchen Fallstudien von erfolgreichen Schulen, vor allem von Schulen, die unter schwierigen Bedingungen, also wider Erwarten, erfolgreich sind. Von deren Praxis können wir eher lernen, was sie konkret besser machen können, als aus dünnen Testergebnissen.

BildungsforscherIn: Akzeptiert man, dass menschliche Erkenntnis immer beschränkt, dass menschliches Handeln immer unzulänglich ist, dann erreichen die neuen Leistungstests ein erfreulich hohes Niveau des Möglichen. Deshalb verdienen sie eine faire Chance – als Versuch und als ein Element in einem größeren Verbund. Denn schon in der IEA-Studie, mehr noch aber in PISA und IGLU nähern wir uns Prüfungsformen, die im schulischen Alltag in der Breite und Differenziertheit oft nicht erreicht werden:

- ein großes Spektrum anspruchsvoller Leistungen;
- offene neben geschlossenen Aufgaben;
- zwei BewerterInnen statt nur einer Person (wie meist im Unterricht).

Ertragreich werden diese Studien aber nur sein, wenn die Programme langfristig angelegt sind, so dass Entwicklungen erkennbar werden. Und nur dann können die ForscherInnen auch aus den ersten Studien lernen und ihre Instrumente kontinuierlich verbessern. Deshalb würde ich jetzt nicht das Land mit einem Testprogramm nach dem anderen überziehen, sondern lieber verschiedene Maßnahmen in überschaubarem Maßstab erproben. So ist auch am ehesten das Vertrauen und die Mitarbeit der Betroffenen zu gewinnen.

Schulaufsicht: Ich stimme dem zu, aber politisch ist das kaum zu machen. Der öffentliche Druck auf die Ministerien ist immens. Eigentlich bräuchten wir ein differenziertes Rechenschaftssystem, in dem Leistungsmessungen einen wichtigen, aber auch nur einen begrenzten Beitrag leisten können. Zudem ist zu bedenken, dass dieser Beitrag unterschiedlich ausfällt, je nachdem, ob sie den Lernerfolg von SchülerInnen, den Lehrerefolg von LehrerInnen oder die Wirksamkeit des Schulsystems insgesamt erfassen sollen. Mir kommt es darauf an, dass Fragen untersucht werden, zu denen es auch tatsächlich Handlungsoptionen gibt. Also müssten die PolitikerInnen ihre Fragen *vorweg* formulieren.

Außerdem finde ich es wichtig, dass die Interpretation der Ergebnisse nicht in der Hand eines einzelnen Forscherteams liegt, sondern dass durch konkurrierende Interpretationen offen gelegt wird, wo die Daten in der Tat eindeutig und wo unterschiedliche Deutungen möglich sind. Es geht doch nicht an, dass die einen aus TIMSS folgern, Japan habe wegen seines problemlösenden Unterrichts so gut abgeschnitten und deshalb müsse eine anspruchsvollere „Lernkultur“ entwickelt werden, während andere aus derselben Studie folgern, Japan habe so gut abgeschnitten, weil die privaten Nachhilfeschoolen, die Jukus, nachmittags durch schlichtes Pauken den Lernerfolg gesichert hätten.

Testprogramme sind keine Selbstläufer, und deshalb kommt der sozialen Kontrolle von Interpretationen und Entscheidungen eine Schlüsselfunktion zu.

ModeratorIn: Ich versuche einen Ausblick. Bei „Evaluation“ geht es um mehr als um Tests. Vorhin war die Rede von der Pflicht zur „Rechenschaft“. Die betrifft doch alle: *SchülerInnen* sollten sich um Klarheit über die eigenen Ziele und über den Erfolg der eigenen Arbeit bemühen und in der Übernahme dieser Verantwortung gefordert und gefördert werden.

LehrerInnen sollten Lernprozesse im Blick auf einzelne SchülerInnen und auf die Klasse insgesamt beobachten und auswerten, den eigenen Unterricht beobachten (lassen) und Erfahrungen im Gespräch mit KollegInnen und Eltern durchdenken, um die Wirkungen der eigenen Arbeit zu überprüfen.

Das *Kollegium* sollte gemeinsam – im Blick auf die Schule als ganze – ein Schulprogramm entwickeln und anhand fokussierter Bestandsaufnahmen (auch mit Hilfe von externen “critical friends”) regelmäßig fortschreiben.

Die *Schulaufsicht* sollte in den schulinternen Diskussionen den Fremdblick stärken, kritische Fragen stellen, Weiterentwicklung fordern, vor allem aber konstruktive Ansätze anerkennen und unterstützen.

Das *Ministerium* sollte mit Blick auf das Gesamtsystem allgemeine Problemstellen ausmachen, politische Prioritäten überprüfen und Ressourcen ziel- und bedarfsorientiert organisieren.

Und noch eins: Vieles, was im Moment als “Evaluation” verkauft wird, ist eigentlich Grundlagenforschung. Zu den Zusammenhängen zwischen verschiedenen Faktoren im System gibt es doch schon eine Fülle von Publikationen und Daten, die man mit vergleichsweise geringerem Aufwand aufbereiten und gezielt auf aktuelle Fragen hin auswerten könnte, statt mit hohen Kosten Berge neuer Daten zu erheben. Wenn von Ökonomie die Rede ist, sollte dieser Anspruch auch für die Evaluation selbst gelten.

Müsste man nicht stattdessen SchülerInnen, Eltern, LehrerInnen und die Öffentlichkeit differenzierter nach ihren Ansprüchen an die Schule, nach ihren unterschiedlichen Erfahrungen und Kritikpunkten befragen? Damit würden die *Wertfragen* und die konkurrierenden Perspektiven auf Unterricht deutlicher, und man könnte den Blick zugleich auf unzulängliche *Bedingungen* schulischer Arbeit erweitern. Es geht doch nicht nur darum, was ist, sondern auch darum, was *sein* soll.